Scientific Journal of Riga Technical University
Computer Science. Information Technology and Management Science
2011
Volume 49

Distance Metrics Selection Validity in Cluster Analysis

Peter Grabusts, Rezekne Higher Educational Institution

***Abstract* – In cluster analysis data are divided into groups according to a specific criterion called metrics. Traditionally the metrics of choice has been Euclidean distance. This article studies other distance metrics used in cluster analysis– Manhattan distance, Cosine distance and Pearson correlation measure. In k-means clustering algorithm these metrics were used to determine cluster centers and the clustering correctness was evaluated. It was found that the clustering results were very similar. The article also contemplates to evaluate clustering validity criteria.**

***Keywords* – clustering algorithms, cluster validity, k-means, metrics**

I. Cluster Analysis Method

Cluster analysis is used to automatically generate a list of patterns by a training set. All the objects of this sample are presented to the system without the indication to which pattern they belong. The cluster analysis is based on the hypothesis of compactness. It means that methods of cluster analysis enable one to divide the objects under investigation into groups of similar objects frequently called clusters or classes. Given a finite set of data X, the problem of clustering in X is to find several cluster centres that can properly characterize relevant classes of X. In classic cluster analysis, these classes are required to form a partition of X such that the degree of association is strong for data within blocks of the partition and weak for data in different blocks.

As a data mining function, clustering can be used as a standalone tool to gain insight into the distribution of data, to observe the characteristics of each cluster, and to focus on a particular set of clusters for further analysis [1]. Clustering is one of the most fundamental issues in data recognition. It plays a significant role in searching for structures in data. It may serve as a pre-processing step for other algorithms, such as classification and characterization, which will operate on the detected clusters [2].

In general, clustering algorithms are used to group some given objects defined by a set of numerical properties in such a way that the objects within a group are more similar than the objects in different groups. Therefore, a particular clustering algorithm needs to be given a criterion to measure the similarity of objects, how to cluster the objects into groups. One of the most widely used k-means clustering algorithm uses the Euclidean distance to measure the similarities between objects. K-means clustering algorithms need to assume that the number of groups (clusters) is known a priori. Table 1 outlines the k-means clustering algorithm [3].

TABLE I
An Outline of the K-means Algorithm

K-means clustering
1. Decide on a value for *k*.
2. Initialize the *k* cluster centers (randomly, if necessary).
3. Decide the class memberships of the *N* objects by assigning them to the nearest cluster center.
4. Re-estimate the *k* cluster centers, by assuming the memberships found above are correct.
5. If none of the *N* objects changed membership in the last iteration, exit. Otherwise go to 3.

The term "cluster analysis" actually comprises a set of different classification algorithms. A common question frequently asked by researchers is: how to organize the data observed into clear structures? A viewpoint exists that unlike many other statistical procedures, methods of cluster analysis are commonly used when the researcher has not got any prior hypotheses regarding classes but is still at the descriptive stage of investigation. It should be noted that the cluster analysis determines the most possible meaningful decision [4]-[6].

Cluster analysis is used to automatically generate a list of patterns by a training set. All the objects of this sample are presented to the system without the indication to which pattern they belong. The cluster analysis is based on the hypothesis of compactness. It means that methods of cluster analysis enable one to divide the objects under investigation into groups of similar objects frequently called clusters or classes. Given a finite set of data X, the problem of clustering in X is to find several cluster centres that can properly characterize relevant classes of X. In classic cluster analysis, these classes are required to form a partition of X such that the degree of association is strong for data within blocks of the partition and weak for data in different blocks.

Similar to other clustering algorithms, k-means clustering has many drawbacks:

- Cluster number, k, must be determined beforehand.
- It is difficult to determine the contribution each attribute makes to the grouping process, since it is assumed that each attribute has the same weight.
- By using the same data, we may never know the real cluster. If the number of data is a few, by inputting data in a different order, a result may be a different cluster.
- In case there are not many numbers of data, the cluster will be significantly determined by the initial grouping.

- Weakness of arithmetic mean is not robust to outliers. As a result, the centroid may be pulled away from the real data by outliers.
- It is sensitive to initial condition, since different initial condition may lead to different result of cluster. The algorithm may be trapped in the local optimum.
- As a result, one gets a circular cluster shape which is based on distance [7].

II. Distance Metrics Overview

Nowadays the concept of regularity or similarity is acquiring more and more attention in the representation of intelligent data processing system operation. In many cases it is necessary to ascertain in what manner the data are interrelated, how various data differ or agree with each other, and what the measure of their comparison is. In various dictionaries the term "regularity" or "similarity" is interpreted as similarity, conformity with a law or conclusion by analogy. Regularity can be considered to be determined correctly if it explains the results of all experiments that relate to the given area of operation.

The main purpose of metrics learning in a specific problem is to learn an appropriate distance/similarity function. Metrics learning has become a popular issue in many learning tasks and can be applied in a wide variety of settings, since many learning problems involve a definite notion of distance or similarity [8]. A metrics or distance function is a function which defines a distance between elements of a set [9], [10]. A set with a metric is called a metric space. In many data retrieval and data mining applications, such as clustering, measuring similarity between objects has become an important part. In general, the task is to define a function Sim(X,Y), where X and Y are two objects or sets of a certain class, and the value of the function represents the degree of "similarity" between the two. Formally, a distance is a function D with nonnegative real values, defined on the Cartesian product X x X of a set X. It is called a metrics on X if for every x,y,z ε X:

- D(x,y)=0 if x=y (the identity axiom);
- D(x,y) + D(y,z) ≥ D(x,z) (the triangle inequality);
- D(x,y)=D(y,x) (the symmetry axiom).

A set X provided with a metric is called a metric space.

Euclidean distance is the most common use of distance – it computes the root of square differences between coordinates of a pair of objects [7]:

$$D_{XY} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2} \qquad (1)$$

Manhattan distance or city block distance represents distance between points in a city road grid. It computes the absolute differences between coordinates of a pair of objects [7]:

$$D_{XY} = \sum_{k=1}^{d} | x_{ik} - x_{jk} | \qquad (2)$$

Minkowski distance is the generalized metric distance:

$$D_{XY} = (\sum_{k=1}^{d} | x_{ik} - x_{jk} |^p)^{1/p} \qquad (3)$$

Note that when p=2, the distance becomes the Euclidean distance. When p=1, it becomes city block distance.

The distance measure can also be derived from the correlation coefficient, such as the Pearson correlation coefficient. Correlation coefficient is standardized angular separation by centering the coordinates to its mean value. It measures similarity rather than distance or dissimilarity [7]:

$$r_{ij} = \frac{\sum_{k=1}^{d} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{d} (x_{ik} - \bar{x}_i)^2 \sum_{k=1}^{d} (x_{jk} - \bar{x}_j)^2}} \qquad (4)$$

where $\bar{x}_i = \frac{1}{d} \sum_{k=1}^{d} x_{ik}$.

Noticing that the correlation coefficient is in the range of [-1, 1], with 1 and -1 indicating the strongest positive and negative correlation respectively, we can define the distance measure as

$$D_{XY} = (1 - r_{ij}) / 2 \qquad (5)$$

When using correlation coefficients for distance measures, it should be taken into consideration that they tend to detect the difference in shapes rather than determining the magnitude of differences between two objects.

Cosine distance is the angular difference between two vectors:

$$D_{XY} = \cos(\theta) = \frac{X \bullet Y}{\|X\| \|Y\|} = \frac{\sum_{i=1}^{n} X_i \times Y_i}{\sqrt{\sum_{i=1}^{n} (X_i)^2} \times \sqrt{\sum_{i=1}^{n} (Y_i)^2}} \qquad (6)$$

The summary of the metrics is shown in Table 2.

TABLE II
DISTANCE MEASURES AND THEIR APPLICATIONS

Measure	Examples and applications
Euclidean distance	K-means with its variants
Manhattan distance	Fuzzy ART, clustering algorithms
Cosine distance	Text Mining, document clustering
Pearson correlation	Widely used as the measure for microarray gene expression data analysis

The purpose of the experimental part was to test the operation of the k-means algorithm by applying different metrics. Four different metrics have been chosen: Euclidean distance, Manhattan distance, Cosine distance and Pearson correlation. In the course of the experiments in order to determine cluster centres in the k-means clustering algorithm, all four metrics have been used sequentially. The results obtained have been analyzed and the clustering correctness has been tested.

During the experiment, the well-known Fisher's IRIS data set was employed [11], containing three species classes of 50 elements each: setosa, versicolor and virginica. Each species has four attributes: SL - sepal length, SW - sepal width, PL - petal length, PW - petal width. This data set is used in cluster analysis, because the data set contains only two clusters with rather obvious separation. One of the clusters contains the Iris setosa species, while the other cluster contains both Iris virginica and Iris versicolor and is not separable without the species information Fisher used.

The experimental part has been carried out in the Matlab environment [12]. The results of the experiments are shown in Table 3.

TABLE III
CLUSTERING RESULTS BY APPLYING DIFFERENT METRICS

Distance	Euclidean	Manhattan	Cosine	Correlation
Cluster centres	50.06 34.28 14.62 2.46 68.50 30.74 57.42 20.71 59.02 27.48 43.94 14.34	50 34 15 2 57 27 42 13 65 30 54 19	0.80 0.55 0.23 0.04 0.75 0.35 0.53 0.16 0.71 0.32 0.59 0.22	0.68 0.24 -0.29 -0.63 0.62 -0.35 0.34 -0.61 0.69 -0.23 0.20 -0.66
Cluster1 contains:	Records from cluster 1 – **50** Records from cluster 2 – 0 Records from cluster 3 – 0	Records from cluster 1 – **50** Records from cluster 2 – 0 Records from cluster 3 – 0	Records from cluster 1 – **50** Records from cluster 2 – 0 Records from cluster 3 – 0	Records from cluster 1 – **50** Records from cluster 2 – 0 Records from cluster 3 – 0
Cluster2 contains:	Records from cluster 1 – 0 Records from cluster 2 – **48** Records from cluster 3 – **2**	Records from cluster 1 – 0 Records from cluster 2 – **39** Records from cluster 3 – **11**	Records from cluster 1 – 0 Records from cluster 2 – **45** Records from cluster 3 – **5**	Records from cluster 1 – 0 Records from cluster 2 – **47** Records from cluster 3 – **3**
Cluster3 contains:	Records from cluster 1 – 0 Records from cluster 2 – **14** Records from cluster 3 – **36**	Records from cluster 1 – 0 Records from cluster 2 – **4** Records from cluster 3 – **46**	Records from cluster 1 – 0 Records from cluster 2 – 0 Records from cluster 3 – **50**	Records from cluster 1 – 0 Records from cluster 2 – **3** Records from cluster 3 – **47**
Correct-ness:	For cluster 1 – **100 %** For cluster 2 - **96 %** For cluster 3 - **72 %**	For cluster 1 – **100 %** For cluster 2 - **78 %** For cluster 3 - **92 %**	For cluster 1 – **100 %** For cluster 2 - **90 %** For cluster 3 - **100 %**	For cluster 1 - **100 %** For cluster 2 - **94 %** For cluster 3 - **94 %**

The above table shows that all metrics correctly recognize cluster 1 records. Cluster 2 records are best recognized by Euclidean distance, whereas cluster 3 records – by Cosine distance. The following figure in the form of a chart shows potentialities of different metrics in clustering (see Fig. 1).

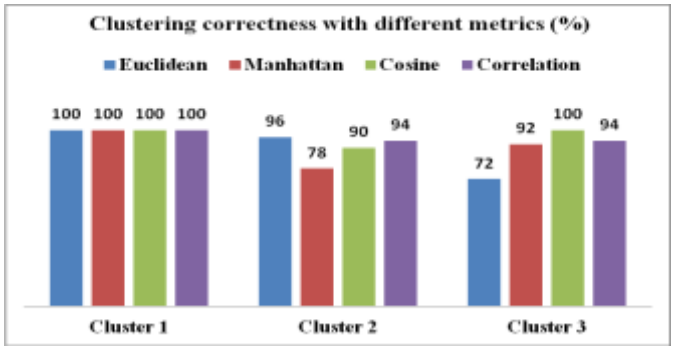

Fig. 1. Clustering correctness

The visualization of clustering may be useful when analyzing results. For data visualization purposes, 2D projections can be used showing the distribution of particular parameters with respect to each other, while dendrogram graphs are normally used for visualization of the formation of clusters (see Fig. 2, Fig. 3, Fig. 4 and Fig. 5).

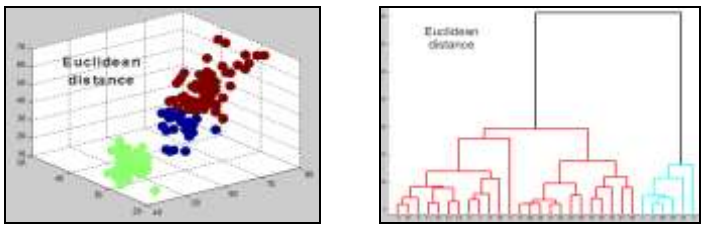

Fig. 2. Clustering results for Euclidean distance

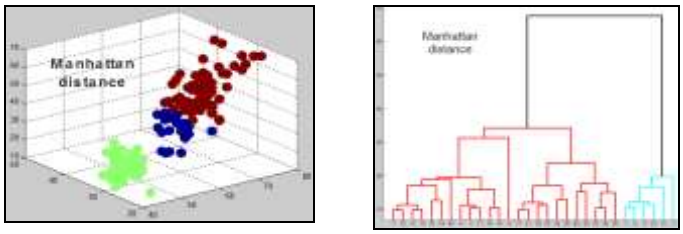

Fig. 3. Clustering results for Manhattan distance

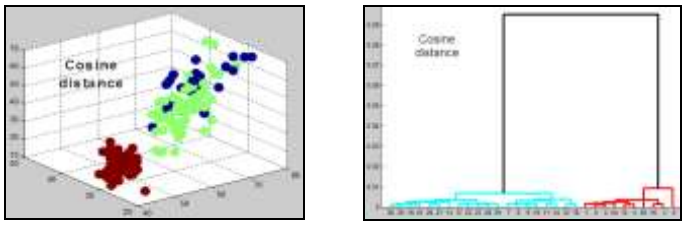

Fig. 4. Clustering results for Cosine distance

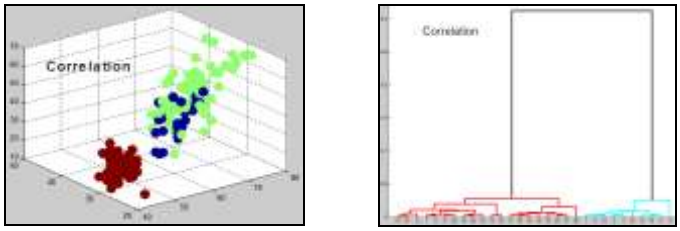

Fig. 5. Clustering results for Correlation measure

Based on the tables and figures it can be concluded that the results obtained by applying all four metrics are very similar. None of the metrics shows dominance that could allow considering it the best metrics. Traditionally Euclidean distance is used in clustering algorithms; however, the choice of other metric in definite cases may be disputable. It depends on the task, the amount of data and on the complexity of the task.

III. Evaluation of Clustering Algorithms

While advancing into the study a question of clustering validity criteria, i.e. determining a numeric criterion to evaluate clustering result, was brought about.

Cluster validity is a method to find a set of clusters that best fits natural partitions (number of clusters) without any class information.

There are three fundamental criteria to investigate the cluster validity: external criteria, internal criteria, and relative criteria [4]. In this case only external cluster validity index was analyzed.

Given a data set X and a clustering structure C derived from the application of a certain clustering algorithm on X, external criteria compare the obtained clustering structure C to a pre-specified structure, which reflects a priori information on the clustering structure of X. For example, an external criterion can be used to examine the match between the cluster labels with the category labels based on a priori information. In contrast to external criteria, internal criteria evaluate the clustering structure exclusively from X, without any external information. For example, an internal criterion would use the proximity matrix of X to assess the validity of C. Relative criteria compare C with other clustering structures, obtained from the application of different clustering algorithms or the same algorithm but with different parameters on X, and determine which one may best represent X in some sense. For example, a relative criterion would compare a set of values of K for the k-means algorithm to find the best fit of the data [4].

In the general approach of cluster validity, the basic idea is to test whether the data points in the data set are randomly structured or not. The test is based on the null hypothesis, H_0, which is the hypothesis of random structure of the data set. If null hypothesis is accepted, then the data in the data set are randomly distributed [6].

Based on the external criteria, there are two different approaches:

- Comparing the proximity matrix Q to the partition P.
- Comparing the resulting clustering structure C to an independent partition of the data P, which was built according to intuition about the clustering structure of the data set [6].

The first approach compares the proximity matrix Q to the partitioning P.

The second approach is more interesting.

If P is a pre-specified partition of data set X with N data points and is independent of the clustering structure C resulting from a clustering algorithm, then the evaluation of C by external criteria is achieved by comparing C to P. Considering a pair of data points xi and xj of X, there are four different cases based on how xi and xj are placed in C and P.

- Case 1: x_i and x_j belong to the same clusters of C and the same category of P.
- Case 2: x_i and x_j belong to the same clusters of C but different categories of P.
- Case 3: x_i and x_j belong to different clusters of C but the same category of P.
- Case 4: x_i and x_j belong to different clusters of C and different category of P.

Correspondingly, the numbers of pairs of points for the four cases are denoted as a, b, c and d (see example in Fig. 6). Because the total number of pairs of points is N(N-1)/2, denoted as M, we have

$$M = a + b + c + d = \frac{n(n-1)}{2}, \qquad (7)$$

where n is the number of data points in the data set.

The data set consists of seven data points. Illustration of four different cases on how a pair of data points is placed in a pre-specified partition P and a resulting clustering structure C is shown in Table 4.

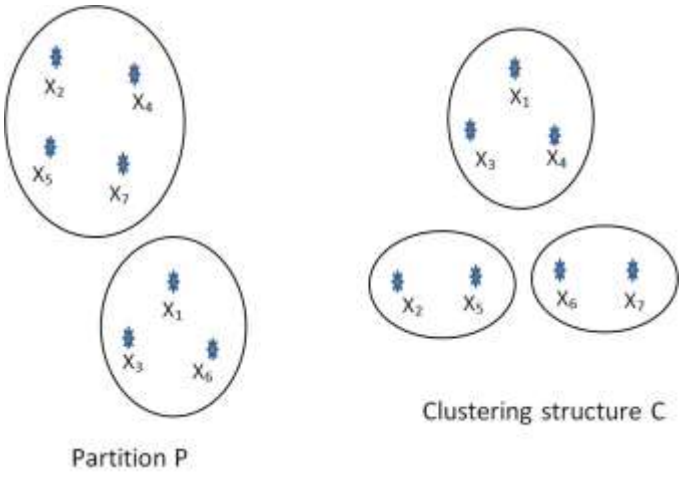

Fig. 6. The numbers of pairs of points for the four cases

TABLE IV
FOUR DIFFERENT CASES OF DATA POINTS

Case	Pairs of data points	Total
a	x_1 and x_3; x_2 and x_5	2
b	x_1 and x_4; x_3 and x_4; x_6 and x_7	3
c	x_1 and x_6; x_2 and x_4; x_2 and x_7; x_3 and x_6; x_4 and x_5; x_4 and x_7; x_5 and x_7	7
d	x_1 and x_2; x_1 and x_5; x_1 and x_7; x_2 and x_3; x_2 and x_6; x_3 and x_5; x_3 and x_7; x_4 and x_6; x_5 and x_6	9

Some commonly used external indices for measuring the match between C and P are as follows:

Rand index:

$$R = \frac{a+d}{M} \tag{8}$$

Jaccard coefficient:

$$J = \frac{a}{a+b+c} \tag{9}$$

Fowlkes and Mallows index:

$$FM = \sqrt{\frac{a}{a+b}\frac{a}{a+c}} \tag{10}$$

Hubert's index:

$$\Gamma = \frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} X_{ij}Y_{ij} \tag{11}$$

The range for the first three indices is [0,1]. High values of these indices indicate great similarity between C and P. High values of the Hubert's index indicate a strong similarity between X and Y; its range is also [0,1]. The major difference between the latter two statistics is that the Rand index emphasizes the situation that pairs of points belong to the same group or different groups in both C and P, but the Jaccard coefficient excludes d in the similarity measure [4].

Rand index suggests an objective criterion for comparing two arbitrary clusterings based on how pairs of data points are clustered. Given two clusterings, for any two data points there are two cases:

The first case is that the two points are placed together in a cluster in each of two clusterings or they are assigned to different clusters in both clusterings.

The second case is that the two points are placed together in a cluster in one clustering and they are assigned to different clusters in the other.

In the experimental part two external validity indices were calculated for IRIS data set – Rand index and Hubert's index (true class labels are known) (see Fig. 7 and Fig. 8).

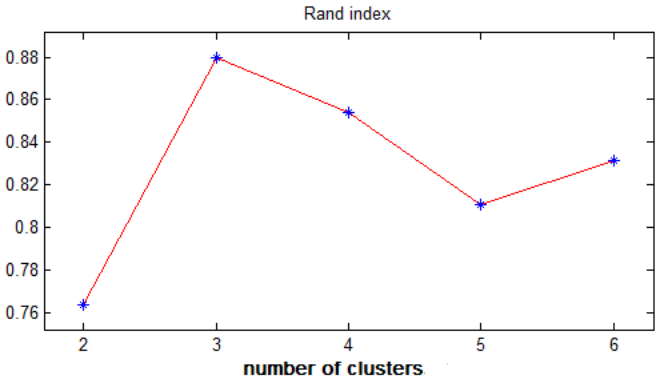

Fig. 7. Rand index

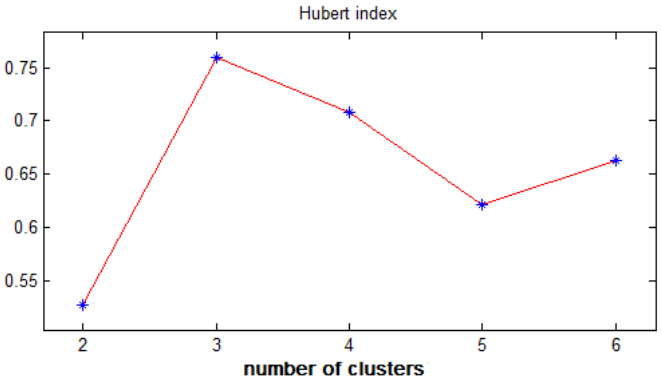

Fig. 8. Hubert's index

Cluster validity criteria are used to examine hierarchical clustering structures, partitional clustering structures or similar. Estimating the number of clusters in unknown data is a main task in cluster validation. It is reasonable to examine the clustering results with different validation methods before making any conclusions [4].

IV. CONCLUSIONS

In the cluster analysis it is necessary to classify data or find relationships in them; therefore the term "relationship" is gaining importance in the context of intellectual data analysis systems. It is often required to determine how the data are related, what the similarity or the difference of different data is, what the metrics is to compare these data. For these purposes one can use different clustering algorithms that divide the data into groups according to specific criteria called metrics; metrics in this context means distance between points belonging to a cluster.

The article inspects the operation of the classical k-means clustering algorithm using various metrics: Euclidean distance, Manhattan distance, Cosine distance and Pearson correlation measure. In the course of experiments cluster centers in the k-means clustering algorithm were calculated using the four aforementioned metrics consecutively. The obtained results were analyzed and the correctness of clustering was scrutinized. Traditionally, clustering algorithms use Euclidean distance but the choice of other metrics can be considered in some cases. It depends on the task that is being solved, the amount and the complexity of data. The study revealed that the clustering results with all examined metrics used were very similar. None of the chosen metrics had a significant predominance that would grant preference and declare it the best metrics. The author considers it valuable to concentrate on cluster validity problems in further research.

REFERENCES

[1] S. Jahirabadkar, P. Kulkarni, ISC- *Intelligent Subspace Clustering, A Density Based Clustering Approach for High Dimensional dataset*, World Academy of Science, Engineering and Technology, 55, 2009.
[2] J. Han M. Kamber, and A. K. H. Tung. *Geographic Data Mining and Knowledge Discovery*, chapter Spatial Clustering Methods in Data Mining: A Survey, pages 1–29. Taylor and Francis, 2001.
[3] B.S. Everitt, *Cluster analysis*. Edward Arnold, London, 1993.
[4] R. Xu and D.C. Wunch, *Clustering*. John Wiley & Sons, 2009, pp. 263-278.
[5] L. Kaufman and P.J. Rousseeuw, Finding groups in data. An introduction to cluster analysis. John Wiley & Sons, 2005.
[6] G. Gan., C. Ma and J. Wu, *Data clustering: Theory, algorithms and applications*. ASA-SIAM series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA, 2007.
[7] P. Grabusts, *The choice of metrics for clustering algorithms.* Proceedings of the 8th International Scientific and Practical Conference. Volume II, "Environment. Technology. Resources", Rezekne, 2011.] Thomson Reuters Web of Science online: http://zdb.ru.lv/conferences/3/
[8] R. Agrawal, C. Faloutsos and A. Swami, *Efficient similarity search in sequence databases*. Proc. 4th Int. Conf. On Foundations of Data Organizations and Algorithms, Chicago. pp. 69-84, 1993.
[9] M. Li, X. Chen, B. Ma and P. Vitanyi, *The similarity metric*. IEEE Transactions on Information Theory, vol.50, No. 12, pp.3250-3264, 2004.
[10] P. Vitanyi, *Universal similarity*. ITW2005, Rotorua, New Zealand, 2005.
[11] R.A. Fisher, The use of multiple measurements in taxonomic problems. Ann. Eugenics, 7(2), p.179-188, 1936.
[12] MathWorks homepage, www.mathworks.com [Accessed: Sept. 30, 2011].

Peter Grabusts was born in Rezekne, Latvia. He received his dr.sc.ing. degree in Information Technology from Riga Technical university in 2006. Since 1996 he has been working at Rezekne Higher Education Institution. Since 2006 he is an Associate Professor at the Department of Computer Science.
His research interests include data mining technologies, neural networks and clustering methods. His current research focuses on techniques for clustering and fuzzy clustering.

Pēteris Grabusts. Attāluma metrikas izvēles pamatotība klasteranalīzē
Klasteranalīzē ir nepieciešams kaut kādā veidā klasificēt datus vai atrast likumsakarības tajos, tāpēc jēdziens „likumsakarība" iegūst arvien lielāku nozīmi intelektuālās datu analīzes kontekstā. Bieži ir nepieciešams noskaidrot – kādā veidā dati ir saistīti savā starpā, kāda ir dažādu datu līdzība vai atšķirība, kāds ir šo datu salīdzināšanas mērs. Tādam nolūkam var izmantot dažādus klasterizācijas algoritmus, kas datus sadala grupās pēc noteiktiem kritērijiem – metrikas. Ar metriku šajā kontekstā tiek saprasta distance (attālums) starp klasterā ietilpstošajiem punktiem. Darbā tika pārbaudīta klasiskā klasterizācijas algoritma *k-means* darbības rezultāti ar dažādām metrikām: Eiklīda distanci, *Manhattan* distanci, *Cosine* distanci un Pīrsona korelācijas koeficientu. Eksperimentu gaitā *k-means* klasterizācijas algoritmā klasteru centru noteikšanai secīgi tika izmantotas minētās četras metrikas. Iegūtie rezultāti tika analizēti un tika pārbaudīts klasterizācijas korektums. Tradicionāli klasterizācijas algoritmos izmanto Eiklīda distanci, taču citas metrikas izvēle atsevišķos gadījumos var būt diskutējama. Tas atkarīgs no risināmā uzdevuma, datu apjoma un sarežģītības. Tika konstatēts, ka klasterizācijas rezultāti visu apskatāmo metriku izmantošanā ir ļoti līdzīgi. Nevienai no izvēlētajām metrikām nebija izšķirīga pārsvara, kas varētu garantēti pasludināt to par labāko. Darba izstrādes laikā aktualizējās jautājums par klasterizācijas kvalitātes kritērijiem, t.i., skaitliska kritērija noteikšanu, lai varētu novērtēt klasterizācijas rezultātu. Klasterizācijas kvalitātes kritēriji tika novērtēti ar Randa indeksu un Huberta indeksu.

Петерис Грабуст. Обоснование выбора метрики расстояния в кластерном анализе
В кластерном анализе необходимо каким-то образом классифицировать данные или найти в них закономерности, поэтому понятие закономерности имеет большое значение в контексте интеллектуальной обработки данных. Часто приходится выяснять – каким образом данные связаны между собой, какова степень сходства или различия между ними, какова мера сравнения этих данных. Для таких целей можно использовать различные алгоритмы кластеризации, которые группируют данные по определенным критериям – метрики. Под метрикой в этом контексте подразумевается расстояние (дистанция) между точками кластера. В статье проверяются результаты работы классического алгоритма кластеризации *k-means* с различными метриками: Эвклидовым расстоянием, Манхэттенской дистанцией, *Cosine* дистанцией и коэффициентом корреляции Пирсона. Во время экспериментов для определения центров кластеров последовательно применялись все четыре упомянутые метрики. Полученные результаты анализировались, и была проверена корректность кластеризации. Традиционно в алгоритмах кластеризации используется Эвклидово расстояние, но в определенных случаях выбор другой метрики может быть целесообразным. Это зависит от решаемой задачи, объема и сложности данных. Было установлено, что при использовании различных метрик результаты кластеризации были очень схожи. Ни одна из рассматриваемых метрик не имела такого перевеса, чтобы определить ее как наилучшую. Во время написания статьи стал актуальным вопрос о критериях качества кластеризации, т.е. определение численного критерия для оценки качества кластеризации. Оценки качества кластеризации были произведены с помощью индексов *Rand* и *Hubert*.